Scientific Research Publishing

Open Journal of Yangtze Gas and Oil, 2018, 3, 1-10
http://www.scirp.org/journal/ojogas
ISSN Online: 2473-1900
ISSN Print: 2473-1889

Research on Realization of Petrophysical Data Mining Based on Big Data Technology

Yu Ding[1,2], Rui Deng[2,3], Chao Zhu[4*]

[1]School of Computer Science, Yangtze University, Jingzhou, China
[2]Key Laboratory of Exploration Technologies for Oil and Gas Resources (Yangtze University), Ministry of Education, Wuhan, China
[3]School of Geophysics and Oil Resource, Yangtze University, Wuhan, China
[4]The Internet and Information Center, Yangtze University, Jingzhou, China
Email: *37275902@qq.com

How to cite this paper: Ding, Y., Deng, R. and Zhu, C. (2018) Research on Realization of Petrophysical Data Mining Based on Big Data Technology. *Open Journal of Yangtze Gas and Oil*, **3**, 1-10.
https://doi.org/10.4236/ojogas.2018.31001

Received: May 24, 2017
Accepted: January 28, 2018
Published: January 31, 2018

Abstract

This paper studied the interpretation method of realization of data mining for large-scale petrophysical data, which took distributed architecture, cloud computing technology and B/S mode referred to big data technology and data mining methods. Based on petrophysical data mining application of K-means clustering analysis, it elaborated the practical significance of application association with big data technology in well logging field, which also provided a scientific reference for logging interpretation work and data analysis and processing method to broaden the application.

Keywords

Big Data Technology, Data Mining, Logging Field Method

1. Introduction

With the increasing scale of oil exploration and the development of engineering field, the application of high-tech logging tools is becoming more and more extensive. The structural, semi-structured and unstructured complex types of oil and gas exploration data are exploded. In this paper, the petrophysical data was taken as the object; Big data technology and data mining methods were used for data analysis and processing, which mines effective and available knowledge to assist routine interpretation of work and to broaden the scientific way to enhance the interpretation of precision. The research allows full play to great potential of logging interpretation for comparative study of geologic laws and oil and gas prediction.

The rapid development of network and computer technology as well as the large-scale use of database technology makes it possible to extract effective information from petrophysical data in more different ways adopted by logging interpretation. Relying on the traditional database query mechanism and mathematical statistical analysis method, it is difficult to satisfy the effective processing of large-scale data. It tends to be that the data contains a lot of valuable information, but it cannot be of efficient use because the data is in an isolated state and cannot be transformed into useful knowledge applied to logging interpretation work. Too much useless information will inevitably lead to the loss of information distance [1] and useful knowledge which is in the "rich information and lack of knowledge" dilemma [2].

2. Analysis of Big Data Mining of Petrophysical Data

2.1. Processing Methods of Big Data

Big data can be taken as the reasons for the basis of the data scale, and it is difficult to use existing software tools and mathematical methods in a reasonable time to achieve the analysis and processing of data which has the features of large scale, complex structure and many types [3].

At present, the amount of rock physical data information gradually increases more and more types, which is consistent with the basic characteristics of big data. With the advantages of cloud computing in data processing performance and the good characteristics of distributed architecture, the existing C/S mode interpretation method is transformed into B/S mode on basis of distributed architecture. Then, the situation, in which processing capacity of the original client single node is insufficient, can be handled through increasing the horizontal scaling of the monomer processing node and the node server in the condition of the rational allocation and the use of system resources. Meanwhile, the on-line method is adopted for the analysis and processing of the petrophysical data which can store the data mining results and analysis process in the server. Interpreters can interpret process documents through querying the server-side to make a more reasonable explanation of the logging data in the unknown area or the same type of geological conditions, which can achieve the change of data sharing from the lower stage (data sharing) to the advanced stage (knowledge sharing).

The essence of big data processing methods can be seen as the development and extension of grid computing and prior distributed computing. The significance of big data processing does not just lie in the amount of data, but in these massive available data resources in which valuable information can be gained quickly and effectively while the available mode can be mined and the purpose of acquiring new knowledge can be achieved.

2.2. Overview of Big Data Technology

2.2.1. Distributed System Architecture

Distributed file system is mainly used to achieve data access of the local under-

lying and the upper-level file system. It is the software system on the basis of the network with a high degree of cohesion and transparency. The distributed system architecture can be considered as the software architecture design that operates in multiple processors. This paper chooses HDFS open source distributed file system to build software operating environment [4].

HDFS system architecture shown in **Figure 1** adopts master/slave architecture, and an HDFS cluster is composed of a Namenode and a number of Datanodes. The Namenode node is used to manage the namespace of the file system and to handle client access to the file. The Datanodenode is used to manage the literacy requests of the storage and processing of the file system clients on its nodes.

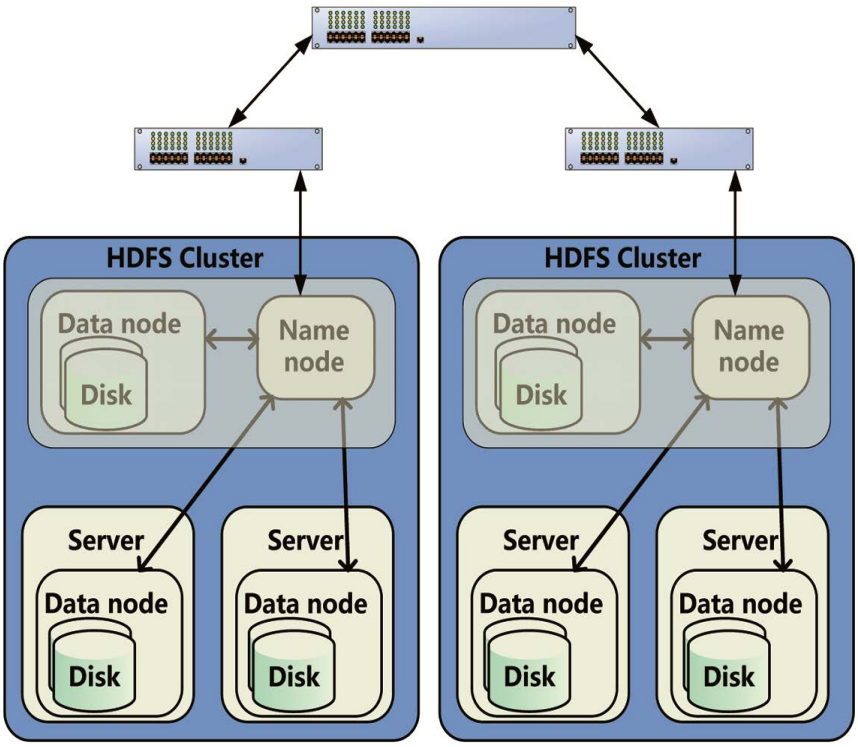


Figure 1. HDFS system architecture.

2.2.2. Cloud Computing Technology

Cloud computing is the Internet-based computing which has been put forward on the basis of the context of the development, being stuck in the bottlenecks, of the traditional computer storage technology and computing capacity (**Figure 2**) [5] [6]. By sharing hardware resources and information to cluster network nodes, large-scale parallel can be achieved and distributed computing to enhance the overall computing power of the system. Combined with the study content of the paper, the cloud computing is applied to the mining of petrophysical data, which can meet the computing requirements of the mining algorithm to solve the problem of insufficient processing capacity of the client nodes in the traditional C/S mode which is the conversion basis of B/S distributed online processing mode.

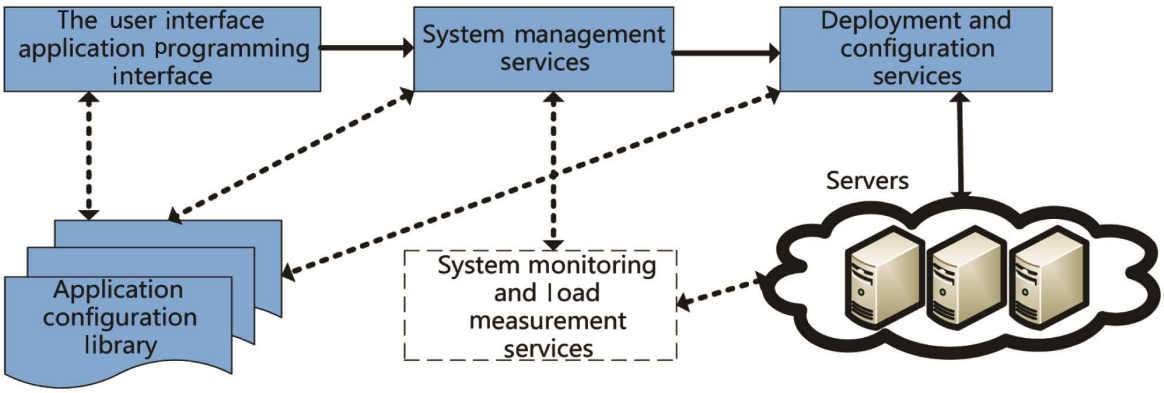


Figure 2. Cloud computing architecture.

2.3. The Combination and Application of Data Mining Methods

2.3.1. Clustering Mining Method

Data clustering is one of the important tasks of data mining. Through clustering, it is possible to clearly identify the regions between inter-class and intra-class of data concentration, which is convenient to understand the global distribution pattern and to discover the correlation between data attributes [7].

In the pattern space S, if given N samples X_1, X_2, ···, X_n, the clustering is defined to find the corresponding regions R_1, R_2, ···, R_m, according to the similarity degree of each other; any of X_i (i = 1, 2, ···, N) is classified into only one instead of the two classes at the same time, to wit, $R_1 \cup R_2 \cup \cdots \cup R_m = R$ and $R_i \cap R_j = \varnothing \quad (i \neq j)$ [8]. Clustering analysis is mainly based on some features of the data set to achieve division according to the specific requirements or rules, which satisfies the following two characteristics under normal circumstances: intra-class similarity, namely, that data items in the cluster should be as similar as possible; inter-class dissimilarity, namely, that data items in the heterogeneous cluster should be as different as possible [9].

2.3.2. Petrophysical Data Clustering Mining Analysis

At present, the analysis and accurate description of sedimentary facies, subfacies and microfacies for favorable reservoir facies zones are an important work in current oilfield exploration and development. The study of sedimentary facies is carried out on the basis of the composition, structure and sedimentary parameters under the guidance of phase pattern and phase sequence. The petrophysical data contains much potential stratigraphic information, and the lithology of the strata often leads to a certain difference in the sampling value of the logging curve. This difference can be seen as the common effects of many factors, such as the lithological mineral composition, its structure and the fluid properties contained in the pores. Because of this, one logging physical value also means some particular lithology of corresponding strata. Coupled with the difference of the formation period and the background, then the combination of the inherent physical characteristics of rock stratum in different geological periods and some

random noise is used to achieve the purpose of lithological and stratigraphic division.

3. Mining Based on K-Means Clustering Analysis

3.1. K-Means Algorithm Principle

Assuming that there is a set of elements, the goal of K-means is to divide the elements of the set into K clusters or classes so that the elements within each cluster have a high degree of similarity while the similarity of elements of different clusters is low, namely, similar elements are clustered into a collection, eventually forming the multiple clustering clustered by feature-similar elements [10].

K-means first randomly generates k objects from n data as the initial clustering center while the rest of the data objects are clustered by calculating the similarity (distance) of each data to the clustering centers (minimum distance between the two points), to divide the data object into the class, and then to recalculate the new cluster of the class center formed by each cluster (cluster the mean of all data objects) to update the cluster class center as the next class center of the iterations. It repeats the clustering process until the criterion function begins to converge.

In this paper, the Euclidean distance is taken as the discriminant condition of similarity measure, and the criterion function E_r is defined as the error sum of the squares of all the data objects to the class center. Obviously, the purpose of the K-means algorithm is to find K divisions of the data set based on the optimal criterion function.

$$E_r = \sum_{i=1}^{K} \sum_{x \in C_i} \left\| x - \overline{x}_i \right\|^2 \tag{1}$$

Here, X represents a data object in the data set; C_i represents the ith cluster, and $\overline{x}_i$ represents the mean of cluster C_i.

3.2. Lithological Division Based on K-Means

The logging physics values of the same layer lithology are relatively stable and generally do not exceed an allowable error. The mean value of the samples in the same layer can be used to represent the overall true value of the similar parts of the surrounding lithology. When the difference between the value of the adjacent sampling point and the mean is within the given error range, the lithological type of the point can be replaced by the lithology corresponding to the mean. Otherwise, it will proceed with the search for the home class until the division of all sampling points is completed. In order to facilitate the study, this paper selects the natural gamma logging curve with strong longitudinal resolution for the division of lithology, while the other passive curves are selected to adjust the division results to improve the accuracy of the decision outcomes in the completion of the lithological division at the same time.

For any two points in the plane (X_1, Y_1) and (X_2, Y_2), the Euclidean distance is

as follows,

$$D_e = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} \quad (2)$$

Here, Figure 3 is taken as the example to show the clustering process of K-means petrophysical data. In Figure 3(a), the black triangles are labeled in two-dimensional space with two-dimensional eigenvectors as coordinates. They can be regarded as examples reflected by two-dimensional data (composed of the data of two logging curves), that is, primitive petrophysical data sets in need of clustering. Three different colored boxes represent the clustering center points (analogical to some lithology) given by random initialization. Figure 3(b) shows the results of the completion of clustering, that is, to achieve the goal of lithological division. Figure 3(c) shows the trajectory of the centroid in the iterative process.

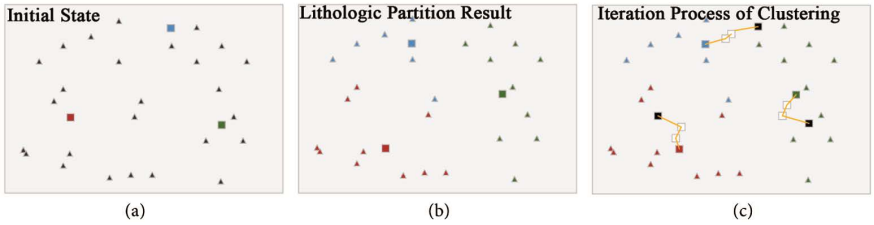


Figure 3. Clustering process of petrophysical data.

The program flow chart is shown in Figure 4 as follow.

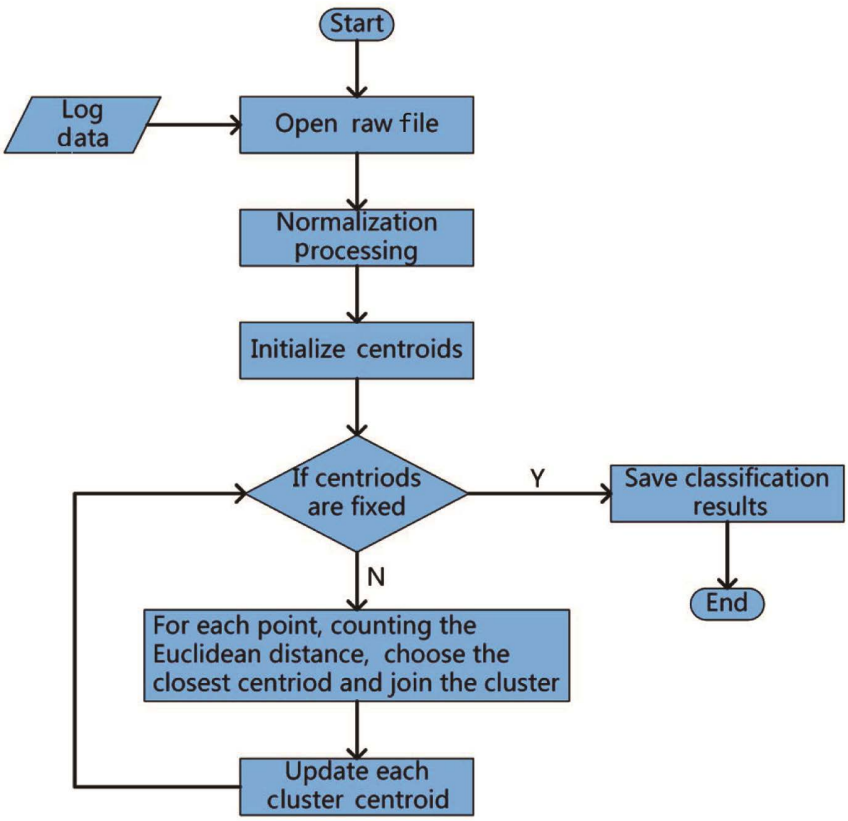


Figure 4. Program flow chart.

3.3. Software Implementation

3.3.1. Distributed Architecture and Cloud Computing Environment

Hadoop operates three modes—the stand-alone, pseudo-distributed and fully distributed. Taking into account the test environment required for the simulation software operation and the main content of this study with the combination of methods to the application, the test environment adopts Hadoop's fully distributed mode in which VMware vSphere 5.5 is used to build another two virtual machines with the CentOS 6 Linux system in the high-performance server equipped with CentOS 6 and the distributed computing is done by three nodes in the cluster (Table 1). Different from the physical node, the cluster node is the use of software virtual composition and the actual operation of the process with differences in performance.

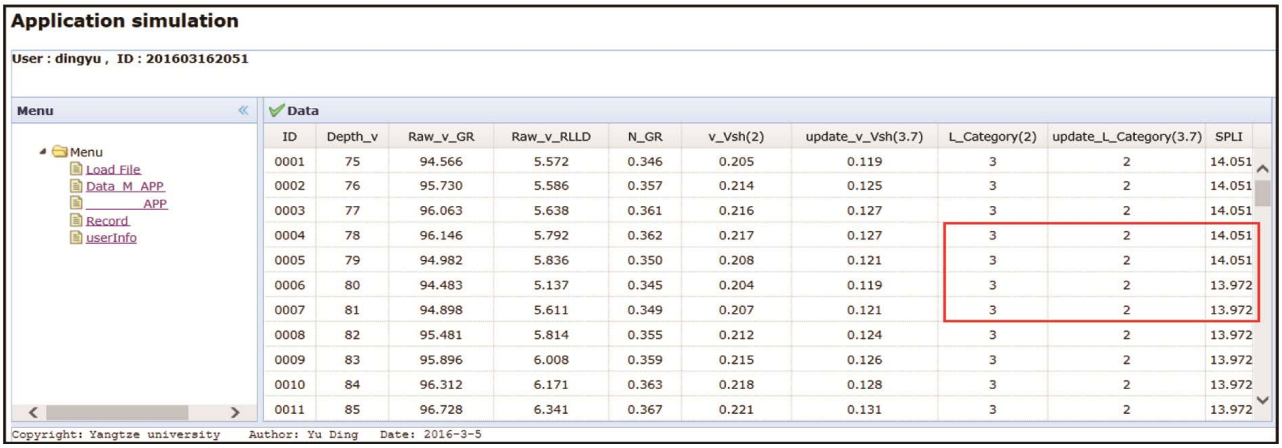
Table 1. Description of the hosts and terminals in the cluster.

Hosttype	Host name	OS	IPaddress	Nodetype
terminal	localhost	Windows 7	10.102.10.35	-
hostmachine	test.com	CentOS 6	10.211.6.1	-
virtualmachine_1	master	CentOS 6	10.211.40.7	master
virtualmachine_2	slave_1	CentOS 6	10.211.40.8	slave
virtualmachine_3	slave_2	CentOS 6	10.211.40.9	slave

3.3.2. Application and Analysis

A total of three production wells in the SZ development Zone of an oilfield are selected to complete the conventional logging interpretation pretreatment by using the collected core material of core section of well walls, relatively complete logging data, geological and drilling data combined with the actual geological conditions, in which the samples with possible existence of the borehole diameter, too large proportion of mud and too high viscosity, leading to measurement curve distortion of the logging instrument, are selected to choose the sample data with the true reflection of the strata information. Then, according to the description of the reservoir performance and the actual division of the corresponding oil and gas standards, the lithology of the working area is divided into four distinct divisions, namely, sandstone, argillaceous sandstone, sandy mudstone and mudstone combined with the core material.

After the K-means algorithm and the lithological judgment condition are programmed, the B/S mode and the cloud computing technology are used to divide the well section lithology of the petrophysical data mining program in the cluster in the built fully distributed simulation environment of Hadoop. The accuracy of lithology is about 78%, and the accuracy of sandstone and mudstone is relatively high which is more than 85%, and the results are shown in Figure 5.



ID	Depth_v	Raw_v_GR	Raw_v_RLLD	N_GR	v_Vsh(2)	update_v_Vsh(3.7)	L_Category(2)	update_L_Category(3.7)	SPLI
0001	75	94.566	5.572	0.346	0.205	0.119	3	2	14.051
0002	76	95.730	5.586	0.357	0.214	0.125	3	2	14.051
0003	77	96.063	5.638	0.361	0.216	0.127	3	2	14.051
0004	78	96.146	5.792	0.362	0.217	0.127	3	2	14.051
0005	79	94.982	5.836	0.350	0.208	0.121	3	2	14.051
0006	80	94.483	5.137	0.345	0.204	0.119	3	2	13.972
0007	81	94.898	5.611	0.349	0.207	0.121	3	2	13.972
0008	82	95.481	5.814	0.355	0.212	0.124	3	2	13.972
0009	83	95.896	6.008	0.359	0.215	0.126	3	2	13.972
0010	84	96.312	6.171	0.363	0.218	0.128	3	2	13.972
0011	85	96.728	6.341	0.367	0.221	0.131	3	2	13.972

Figure 5. Data mining result.

The data in Figure 5 shows that the same notions of SPLI values indicate that they are of the same layer, *i.e.*, lithological consistency or similarity viewed from the results of artificial stratification. Compared with the data mining results, due to the difference of the value of the empirical coefficient in the stratigraphic age, the division results based on the discriminant conditions are different in some logging sampling points. According to the correction process of the core data, the result is related to the value of the empirical coefficient. For a certain section of stratum, the value of 2 may have a relatively high degree of coincidence. Similarly, the value of 3.7 of some layers have a relatively high degree of coincidence. This also shows that the general selection of the single empirical coefficient may have an impact on the accuracy of the interpretation results. On the one hand, viewed form the upper and lower adjacent types, the results are identified correctly which belong to the same kind of lithology. On the other hand, viewed from the comparison of the results of the left and right division, the lithological division has changed while the corresponding SPLI value is not exactly the same, indicating that the data have the value of further fine study.

Therefore, in-depth study of inconsistent results of lithological division can help to find valuable unexpected pattern in petrophysical data. Compared with the experimental methods and empirical methods, this method of "Data to talk", from which the potential correlation and extract knowledge are explored, leads to the more objection and science for some regional empirical parameter values and empirical formulas in scientific induction and summary.

Figure 6 shows the time consumed by executing the same program in a stand-alone node and in a distributed environment, which is 616,273 ms and 282,697 ms respectively. Here, the optimization of the algorithm, compiler selection and hardware device performance differences and other factors are considered comprehensively, and only described from the qualitative perspective, the use of distributed computing can reduce or significantly reduce the time spent on large-scale data processing to improve the overall performance of the system to a certain extent, thus the feasibility of using the big data technology to realize the petrophysical data mining is verified.

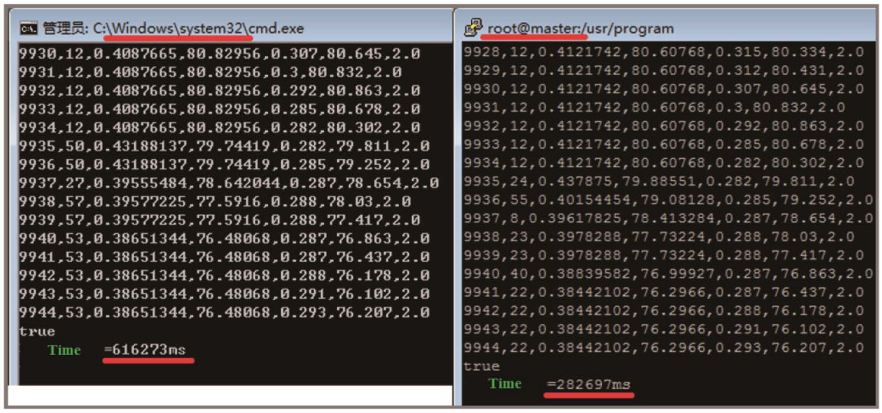


Figure 6. Running time of program in Windows and HDFS.

4. Conclusions

1) The advantages of distributed architecture and cloud computing are used to improve the overall processing capacity of the system, and in the process of large-scale petrophysical data processing, the B/S mode is integrated to achieve data mining to combine big data analysis and processing mechanism with conventional interpretation. The exploratory research idea of the new method of logging interpretation is put forward, with the starting point of discovering the novel knowledge, to provide a scientific reference for the routine widening applied by the interpretation work and data analysis methods.

2) The combination of multidisciplinary knowledge and the rational application of cross technology can perfect the deficiencies in the existing logging interpretation to a certain extent, making the interpretation work of qualitative analysis and quantitative computing of logging data more scientific, with favorable theory and practical guidance significance.

Acknowledgements

This work is supported by Yangtze University Open Fund Project of key laboratory of exploration technologies for oil and gas resources of ministry of education (K2016-14).

References

[1] Wang, H.C. (2006) DIT and Information. Science Press, Beijing.

[2] Wang, L.W. (2008) The Summarization of Present Situation of Data Mining Research. *Library and Information*, **5**, 41-46.

[3] Pan, H.P., Zhao, Y.G. and Niu, Y.X. (2010) The Conventional Well Logging Database of CCSD. *Chinese Journal of Engineering Geophysics*, **7**, 525-528.

[4] Ghemawat, S., Gobioff, H. and Leung, S.-T. (2003) The Google File System. *ACM SIGOPS Operating Systems Review*, **37**, 29-43. https://doi.org/10.1145/1165389.945450

[5] Sakr, S., Liu, A., Batista, D.M., *et al*. (2011) A Survey of Large Scale Data Management Approaches in Cloud Environments. *IEEE Communications Surveys & Tuto-*

rials, **13**, 311-336. https://doi.org/10.1109/SURV.2011.032211.00087

[6] Low, Y., Bickson, D., Gonzalez, J., *et al*. (2012) Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud. *Proceedings of the VLDB Endowment*, **5**, 716-727. https://doi.org/10.14778/2212351.2212354

[7] Song, Y., Chen, H.W. and Zhang, X.H. (2007) Short Term Electric Load Forecasting Model Integrating Multi Intelligent Computing Approach. *Computer Engineering and Application*, **43**, 185-188.

[8] Abraham, B. and Ledolter, J. (1983) Statistical Methods for Forecasting. John Wiley & Sons, Inc., NewJersey.

[9] Farnstrom, F., Lewis, J. and Elkan, C. (2000) Scalability for Clustering Algorithms Revisited. *AcmSigkdd Explorations Newsletter*, **2**, 51-57. https://doi.org/10.1145/360402.360419

[10] Rose, K., Gurewitz, E. and Fox, G.C. (1990) A Deterministic Annealing Approach to Clustering. *Information Theory*, **11**, 373.